Combining Social and Semantic Metadata for Search in a Document Repository

Christoph Herzog[1], Michael Luger[2], and Marcus Herzog[3]

[1] E-Commerce Competence Center, EC3
christoph.herzog@ec3.at
[2] DERI Innsbruck
michael.luger@deri.org
[3] Vienna University of Technology
herzog@dbai.tuwien.ac.at

Abstract. With the success of social applications like Flickr, del.icio.us and YouTube, social software has become the focus of several research initiatives. Especially the idea of combining social and semantic aspects has been recently gaining significant attention in the semantic web community. In this paper we research and discuss the properties of metadata in social systems (folksonomies) compared to metadata in semantic systems (ontologies). We then present our idea for creating a link between a folksonomy and an ontology in order to combine the usability and flexibility of folksonomies with the precision of ontologies for a semantic search application. Our approach is motivated by the requirements of the OnTourism project, which has the goal of creatng a document repository which benefits from both ontology and folksonomy metadata.

1 Introduction

"Web 2.0", a term coined by Tim O'Reilly, has become a much debated topic in the web community. O'Reilly defines "Web 2.0" as platform of (web-based) software that incorporates user participation and "gets better the more people use it" [13]. With the remarkable success of social applications like YouTube, this idea has gained significant attention from the scientific community.

The strength of social applications is that they are easy to use and can generate massive amounts of metadata through an implicit community effort. However, these metadata are semantically not clearly defined and not suitable for reasoning or similar tasks. In this paper we explore the properties of social and semantic metadata. We also present our ideas for how to find relations between the two, based on their observable use in describing documents. Our work is motivated by the OnTourism project, which has the goal of creating a document repository that makes use of both semantic and social metadata.

The paper is organised as follows: In section 2 we review social and semantic metadata and give a comparison. In section 3 we present the OnTourism project and our ideas for a social semantic document repository. This is followed by a review of related work in section 4 and, finally, section 5 concludes the paper.

2 Comparing Social and Semantic Metadata

Folksonomies With the rising popularity of applications like *Flickr*[4], *del.icio.us*[5] and *YouTube*[6], social software has become a research topic. We think the two most important reasons for the high user acceptance of social software are:

1. *Low entry barriers* – Successful social tools like *Flickr* make it as easy as possible for the user to participate. Time, effort and cognitive cost required to use the system are minimised [9].
2. *Instant and delayed gratification* – The tools we examined exhibit patterns of what Ohmukai et al describe as *instant gratification* and *delayed gratification* [12]. Instant gratification is the direct and egoistic benefit users draw from using the system (i.e., organising their photos or bookmarks). Delayed gratification is the added value generated by the community.

Social software places an emphasis on users assigning freely chosen keywords to shared objects. While this is not a new idea, the novel aspect is that not only the author but, to a varying extent, also other users can assign keywords to an object. The tags (keywords) applied by users constitute an emergent vocabulary for which the term *Folksonomy* has been coined by Thomas Vander Wal [15].

A folksonomy in this sense is a set of terms, which from a mathematical point of view can be seen as a tripartite graph with hyper-edges, consisting of *(user, tag, object)* triples. The distribution of the tags follows a power law curve as Vander Wal points out in [16]. Figure 1 shows the tag distribution of a sample of approximately 200.000 tag usages from the del.icio.us folksonomy. The horizontal axis represents the approximately 1000 tags which are used at least 20 times. The vertical axis shows how often each tag is used.

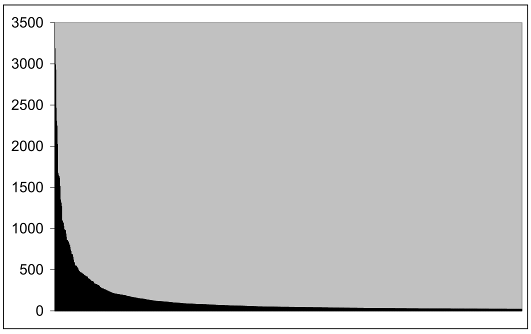

Fig. 1. Tag distribution of a sample from the del.icio.us folksonomy.

The direct benefit of folksonomies lies in making meaningful metadata available in an implicit community effort. This is especially true for systems like photo

[4] A web application for sharing photos. See http://www.flickr.com/

[5] A web-based social bookmarking tool. See http://del.icio.us/

[6] A web application for sharing videos. See http://www.youtube.com/

or video platforms, which are not amenable for text search methods. While the applied tags do not yield explicit semantics, they have the inherent benefit of "speaking the user's language", which is hard to accomplish by top-down ontology engineering. Furthermore, folksonomies conserve minority expressions. When regarding the power law distribution, the most common (*"strong"*) tags represent the major "desire lines" of the emergent vocabulary. However, the long tail of seldom used tags can contain highly specific terms, making the tagged objects amenable to being found by more unusual expressions.

Ontologies The semantic web initiative pursues the goal of creating data and metadata in such a way that not only humans but also machines can make use of it. The idea is that the *meaning* of the data should be expressed in a format which enables it to be processed by computers. Towards this goal, most systems make use of ontologies to describe their data or metadata. An ontology is a model of a real-world domain. This model specifies the most important concepts of that domain, their attributes and relations between concepts.

A particular usage of ontologies found in many semantic systems is the task of inferring new knowledge from facts and rules expressed in an ontology language. Another common task is the execution of search queries on data represented in an ontology language to retrieve semantically meaningful search results. The combination of both leads to semantic search applications that make full use of ontologies in order to provide complete and relevant answers to user queries.

Comparison Folksonomies and ontologies are targeted towards very different applications. In a direct comparison (see table 1), it becomes apparent that ontologies are more suitable for situations where a precise description of data is required and the cost of metadata creation is not an issue. Folksonomies on the other hand perform better when large quantities of metadata are required, where the metadata's precision is not of predominant importance.

3 OnTourism

Our idea of combining folksonomy and ontology metadata is motivated by our current work on the OnTourism project. The main goal of OnTourism is to implement a semantic search functionality on a call centre's existing document repository. The repository contains MS Word and PDF documents which are created by the call centre agents. The expected value of semantic search in this environment is that customer requests can be answered more quickly and more precisely. However, the intended users of the system are not experts in using ontologies, therefore, a social approach will be utilized in order to complement the semantic search function and make the ontology more accessible for the users.

Two kinds of metadata will be employed to desctibe the document's contents and target audience – free user selected keywords (tags) on the one hand and entities of a defined ontology on the other hand. In order to combine the advantages of the ontology and folksonomy metadata, statistical methods will be used

	Folksonomy	Ontology
Structure	flat	hierarchical structure
Creation	by users during the act of using the system	by ontology experts at a given time
Synonyms	no synonym control	synonym control possible
Precision	low precision	high precision
Flexibility	high flexibility	low flexibility
Creation Cost	low, created by users	high, created by experts
Change	highly dynamic, changes constantly	rigid, often has to be recreated to accommodate change
Usability	no expertise required	requires proficiency in handling
Vocabulary	users vocabulary	experts vocabulary
Scalability	works better in a large scale	works better in a small scale

Table 1. Comparison between folksonomies and ontologies (based on [1])

to find probable relations between folksonomy tags and ontology elements. Also, the search functionality itself will be a combination of the results of a semantic search utilising ontology reasoning and a search on the folksonomy tags.

3.1 User Driven Metadata – The OnTourism Folksonomy

Users of the system can add tags through a social bookmarking system. The incentive for adding tags is that the bookmarked documents can be found by searching for the assigned tags, which are the ones that for the user best describe the document. This contributes to the idea of "keeping found things found", i.e., being able to quickly re-find documents once discovered.

The main problem we face is the low number of users. In the call centre, approximately 15 people work with the document repository, limiting the user base. Through the annotation by the document's author, we make sure that each document is tagged at least by one person. However, in the call centre's set-up we must assume that many documents will be tagged by the author only and that even more popular documents receive tags from five or less users.

Another problem folksonomies as flat collections of terms face, is synonym control. People can use different tags with the same meaning (e.g., "Apple" vs. "Mac" vs. "Macintosh"). We do not seek a solution for synonym control in folksonomies, but rather intend to provide strong feedback to the user at the time of entering tags. When entering a keyword, the user will be given suggestions of likely matching or likely related folksonomy terms in real time. In this way we hope to achieve a quick convergence of the vocabulary.

3.2 Structured Metadata – The OnTourism Ontology

In addition to the approach of enriching the application with a social bookmarking functionality, a metadata ontology allows for the annotation of the documents on a more sophisticated and less ambiguous level. This ontology is aimed towards capturing the concepts and relations of the tourism domain as precisely and completely as required for our scenario. A semantic query engine enables to extract more accurate information than regular query engines that solely rely on information retrieval on a purely syntactical level or on unstructured tags.

The actual structure of the ontology is designed towards enabling the creation of an easy to use user interface, both for the process of annotation and for the semantic search component. Within the OnTourism project a close collaboration with Österreich Werbung[7] makes it possible to build the data model based on expert domain knowledge. Initially, the ontology broadly covers the tourism domain in low depth and the domains of special importance to the call centre application in more detail. A basic vocabulary for spatial-location related information such as the Basic Geo Vocabulary[8] is being considered to be incorporated into the ontology. One of the goals of this approach is to use the system's reasoning facilities in order to improve the output of location-related queries.

3.3 The Link between Social and Semantic

In the OnTourism system both ontology and folksonomy metadata will be incorporated. The ontology metadata provides the benefit of enabling a semantic search engine to find precise results and to apply reasoning procedures on the metadata. The folksonomy metadata provides the benefit of generating metadata in terms of a user-driven emergent vocabulary. Our goal, however, is to find relations between the folksonomy and the ontology metadata in such a way that in the overall system the strengths of both are emphasised.

We do not intend to combine the folksonomy and ontology metadata directly, but instead utilise the statistical relation between folksonomy tags and ontology elements, eventually using the folksonomy to enhance the usability of the ontology. Furthermore, semantic search results and the results of a search for tags will be merged into a combined search result.

From the analysis of the co-occurrence between folksonomy terms and ontology elements on single objects, we obtain a "statistical mapping" between the two types of metadata. The goal in extracting these relations is to find for any given tag from the folksonomy the most likely related entities from the ontology.

The user interface for annotating documents will utilise the folksonomy in order to help the user to find desired ontology elements. When entering keywords, the user is presented with suggestions for already existing tags. Once a user chooses such a tag, suggestions are presented for ontology elements most likely related to the keyword by the "mapping" discussed above. An ontology browser will enable the user to select ontology elements not suggsted yet.

Ontology elements which the system suggests may actually be only weakly related to the selected tag, but any useful suggestion improves the usability of the ontology. Moreover, if the suggested elements are not semantically related

[7] The Austrian national tourism organisation.

[8] Basic Geo (WGS84 lat/long) Vocabulary, http://www.w3.org/2003/01/geo/

then they may still be thematically related. In this way the suggestions may be useful for the user as a recommendation (e.g., a user who entered the tag "skiing" might also want to add the ontology element labelled "Mountain").

Mapping Folksonomy to Ontology The algorithm for relating ontology elements to folksonomy tags will be one of the main outputs of the OnTourism project. We shortly present some first ideas for this algorithm.

A standard relatedness measure like the jaccard coefficient [11] or cosine similarity [2] can be selected to define the relatedness between tags based on their co-occurence on documents. A similarity between ontology elements and folksonomy elements will be constructed equivalently.

However, considering the network of related tags from the folksonomy, connected by edges weighted with a relatedness measure as described above, we can extract a hierarchy of clusters and sub-clusters using network analysis methods. For example a specific approach towards this goal is described in [6]. For each cluster of tags in this hierarchy we compute the relatedness to a given ontology element as the sum of the relatedness to the individual tags in a given cluster.

Having done so, we can refine the list of related ontology elements for a given tag by going up along the hierarchy of clusters, adding the ontology elements related to the (bigger) parent cluster with a decreasing weight. In this way ontology elements directly related have a higher weight than ontology elements related by the increasingly fuzzy clusters. Through this procedure, more ontology elements are added to the list of candidates to be suggested to the user. This is especially useful if only few ontology elements co-occur with a given tag.

Those suggested ontology elements actually selected by the user are very likely to have a strong relation to the tag entered by the user. Therefore, the user's choice should strengthen the statistical relation between the tag and the selected ontology element. We are currently investigating how this is best achieved.

3.4 Social Semantic Search

In the OnTourism system, searching for documents will be a combination of semantic search, search on folksonomy terms and full text search. The three search methods can be executed in parallel, with the complete search result being a weighted combination of the three separate (possibly empty) result sets.

Semantic Search — As described above, the semantic search application will provide a graphical interface that allows the user to select concrete objects and attributes from the underlying semantic data model. The actual parameters entered through this interface are then translated into the corresponding formal query, which is then performed on top of the ontology storage component. This semantic search is expected to yield results of a precision not achievable by performing the query upon potentially ambiguous tags.

Folksonomy Search — The second component is the search for documents annotated with specific tags. Where semantic search may fail to retrieve some relevant documents by being too restrictive, searching for folksonomy terms can provide more generous results while still being relevant for the annotated document. Search and ranking in the folksonomy metadata will be based on the *FolkRank* algorithm introduced in [5]. The results of the search on folksonomy metadata will have a lower weight than those of the semantic search.

Full Text Search — The results of full text search will also be considered in order to make sparsely annotated documents retrievable. The weight of the full text search will be considerably smaller than that of the other two methods.

4 Related Work

Several works are being investigated towards the goal of achieving a synergy between social and semantic applications. These works mainly follow one of two approaches [14]: adding more precise semantics to social systems or using a community of users to enhance semantic software.

Examples for adding more semantics to social systems include the idea of semantic enrichment of tags in weblogs [4] or, more generally, the idea to extend the *(object, tag)* graph of a folksonomy towards an *(object, ontology node, tag)* graph [7]. Examples for attempts to add some of the benefit of folksonomies to ontologies include ideas to extend ontologies in a folksonomy-like approach [3] or to add multiple labels to ontology nodes, an idea formulated by Maedche [8].

Another line of works is concerned with extracting semantic relations from folksonomies. While the extraction of complete ontologies from folksonomies appears to be a rather less explored area, there are several works towards extracting at least basic taxonomies from folksonomies [6, 10, 17].

5 Conclusion

Social and semantic software each have their own strengths and weaknesses. Social software is based on a low effort for the individual user in participating. Such systems can generate massive amounts of meaningful metadata. These metadata, however, are neither structured nor controlled. Ontology-based metadata on the other hand has a clear semantic meaning. Creating such semantically rich metadata, however, is expensive in terms of the annotation effort.

In this paper we presented our idea of social semantic document repository, motivated by the requirements of the OnTourism project. In this repository, documents will be annotated with both user defined keywords from an emergent vocabulary (folksonomy) and with metadata from an ontology's vocabulary.

We intend to find relations between the tags and the ontology elements, identified through correlations of the two kinds of metadata when used to annotate single documents. We will use these relations in order to make the ontology more accessible for the users through an appropriate user interface.

Furthermore, in order to search for documents, we will combine the results from search on folksonomy metadata and from a semantic search engine. In this way the search application benefits from both the precision of the ontology and the flexibility of the folksonomy generated by the social component.

References

1. Christine Albrecht. Folksonomy. Master's thesis, Vienna University of Technology, Vienna, Austria, March 2006.
2. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, June 1999.
3. Scott Bateman, Christopher Brooks, and Gord McCalla. Collaborative tagging approaches for ontological metadata in adaptive e-learning systems. In *Proceedings of the 4th International Workshop on Applications of Semantic Web Technologies for E-Learning (SWEL'06)*, June 2006.
4. Steve Cayzer. What next for semantic blogging? In *Proceedings of the SEMANTICS 2006 conference*, pages 71–81, Vienna, Austria, November 2006.
5. Andreas Hotho, Robert Jschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European Semantic Web Conference (ESWC2006)*, pages 411–426, Budva, Montenegro, 2006.
6. Renaud Lambiotte and Marcel Ausloos. Collaborative tagging as a tripartite network. *ArXiv Computer Science e-prints*, December 2005.
7. K. Faith Lawrence and M. C. Schraefel. Freedom and restraint: Tags, vocabularies and ontologies. In *Proceedings of the 2nd IEEE International Conference on Information & Communication Technologies*, Damascus, Syria, 2006.
8. Alexander Maedche. Emergent semantics for ontologies – support by an explicit lexical layer and ontology learning. *IEEE Intelligent Systems - Trends & Controversies*, pages 78–86, February 2002.
9. Adam Mathes. Folksonomies – cooperative classification and communication through shared metadata. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html, December 2004.
10. Peter Mika. Ontologies are us: A unified model of social networks and semantics. In *Proceedings of the International Semantic Web Conference 2005 (ISWC'05)*, pages 522–536, Galway, Ireland, November 2005.
11. Junichiro Mori, Yutaka Matsuo, Mitsuru Ishizuka, and Boi Faltings. Keyword extraction from the web for person metadata annotation. In *Proceedings of the 4th International Workshop on Knowledge Markup and Semantic Annotation, (ISWC2004)*, pages 51–60, Hiroshima, Japan, November 2004.
12. Ikki Ohmukai, Masahiro Hamasaki, and Hideaki Takeda. A proposal of community-based folksonomy with rdf metadata. In *Proceedings of the Workshop on End User Semantic Web Interaction, (ISWC2005)*, Galeway, Ireland, November 2005.
13. Tim O'Reilly. Web 2.0: Compact definition? http://radar.oreilly.com/archives/-2005/10/web_20_compact_definition.html, October 2005.
14. Sebastian Schaffert. Semantic social software: Semantically enabled social software or socially enabled semantic web? In *Proceedings of the SEMANTICS 2006 conference*, pages 99–112, Vienna, Austria, November 2006. OCG.
15. Gene Smith. Folksonomy: Social classification. http://atomiq.org/archives/2004/-08/folksonomy_social_classification.html, August 2004.
16. Thomas Vander Wal. Explaining and showing broad and narrow folksonomies. http://www.personalinfocloud.com/2005/02/explaining_and_.html, February 2005.
17. Xian Wu, Lei Zhang, and Yong Yu. Exploring social annotations for the semantic web. In *Proceedings of the 15th International Conference on World Wide Web (WWW'06)*, pages 417–426, Edinburgh, Scotland, May 2006. ACM Press.